Using Information on Lexical Stress for Utterance Verification

Gies Bouwman, Lou Boves

A^{2}RT, Department of Language and Speech
University of Nijmegen, The Netherlands
{G.Bouwman,L.Boves}@let.kun.nl
http://lands.let.kun.nl/

Abstract

ASR applications like nationwide telephone directory assistance (DA) face the challenge of making a correct classification with only minimal amounts of acoustic data. For this reason, current systems still make too many errors in order to be useful. In the perspective of the idea that 'no recognition' is better than 'misrecognition', a feasible system should therefore detect and reject the least reliable hypotheses. This process is known as utterance verification.

Against the disadvantage of having few information, there is the advantage that isolated utterances have a relatively small degree of prosodic variation, for instance in intonation, speech rate and accent. In this paper we investigate how one can capitalise on this advantage in terms of better utterance verification. We define a number of confidence measures (CMs) on prosodic features and evaluate several linear combinations of one or more CMs.

Experimental results on a field corpus of city names show that a relative improvement of 11.0% Confidence Error Rate can be achieved when compared to a 'conventional' system with only a Log Likelihood Ratio CM.

1. Introduction

In the framework of the EC-funded project SMADA (Speech-driven Multimodal Automatic Directory Assistance), we are investigating the feasibility of adopting ASR technology in a service for nation-wide Directory Assistance (DA). To make automation feasible, callers are guided through a dialogue in which they are first prompted to say the type of listing (residential or business), next to give the city name and then to say the name of a person or a business (or service). If the information so obtained is not enough to identify a unique listing, callers are additionally prompted for the address. In tasks like these, reliable rejection of keywords that were incorrectly recognised is absolutely essential. Recognition errors can be due to confusions between in-vocabulary words, perhaps due to acoustic background noise, but also to Out of Vocabulary (OoV) speech produced by users who do not produce the kind of isolated utterances that the dialogue intends to elicit.

Most existing rejection strategies use confidence measures based on N-best lists (e.g. [1]), Log Likelihood Ratios (LLRs, [2]) or combinations of these ([3]). Recently, the range of strategies has been extended with approaches that make use of knowledge about prosodic features, like phone(me) duration [4][5]. In [5] data are reported that compare phone durations in correct and incorrect word hypotheses. It was observed that phone segments in incorrect word hypotheses often are either too long or too short. The knowledge inferred from these observations was successfully deployed in a new rejection procedure. Parametric models of phone duration in the training material were used to develop confidence measures that tend to reject word hypotheses with phones that deviate too much from the average length.

In this work we seek to extend previous research on confidence measures by combining LLR based measures, and duration based measures with additional prosodic information. In particular, we will examine the possibility to use information about lexical stress of content words to improve the performance of LLR and duration based measures in a large vocabulary, high perplexity isolated utterance recognition task, viz. city name recognition in an automated DA service. In [6] it was shown that lexical stress, despite its uncontested status in phonology, has little to offer in continuous speech recognition tasks, because too many factors intervene between the abstract lexical representation and the actual phonetic realisation of the stressed syllables. However, in isolated utterance recognition the correspondence between lexical stress and (realised) accents may be less involved. Consequently, information about lexical stress might turn out to be a usful feature in the recognition of short utterances.

In this paper we derive and test confidence measures based on phone and syllable duration statistics, and a straightforward way to combine these measures with conventional confidence measures. To this end we developed simple ways for normalising duration for overall speaking rate estimated from short utterances, taking account of the difference between stressed and unstressed syllables. We test the hypothesis that this information can be used to improve duration based rejection methods, despite the fact that estimates of average speech rate may not be very precise in short utterances.

In addition to duration differences, the spectral shape of phones (especially vowels and sonorants) in stressed and unstressed syllables differs in a systematic manner. This knowledge can be used to train separate (anti-)models for phones in stressed and unstressed syllables. We test the hypothesis that using separate models for phones in stressed and unstressed syllables improves the performance of LLR confidence measures.

The remainder of this paper is organised as follows. In the next section, we will introduce a number of confidence measures that are inspired on prosodic knowledge. In Section 3 we provide detailed information about the remaining aspects of our material, software, models and evaluation metrics. Section 4 outlines our experimental setup. Our interpretation of the results presented in Section 5, is part of the discussion in Section 6. Finally, Section 7 summarizes our ideas and findings in the conclusion.

2. Method

2.1. Lexical stress

Although the search strategies of most current ASR decoders are much stronger guided by the emission probabilities of the acoustic models than transition probabilities, the phone duration encodes a considerable amount of information, as illustrated for vowels by Table 1. The information in this table was obtained from our train material, which will be described in more detail in section 3. To compute average phone durations we performed a forced alignment between the phonetic transcription and speech signals in the training corpus. Next we averaged all token durations per phoneme type. Since information about duration is available during recognition, one could use this information to compute a confidence measure for the recognition result.

As can be seen in Table 1, the phoneme's identity is not the only influential factor for duration; lexical stress also is, especially for the long vowels and diphthongs. Long vowels in stressed syllables have significantly longer duration than their unstressed counterparts. The information in Table 1 corroborates the findings reported in [6].

phoneme (SAMPA)	duration (#frames)	
	-stress	+stress
I	9.70	10.84
U	10.91	12.01
Y	10.82	13.37
e:	13.24	14.99
2:	14.54	16.67
a:	13.57	16.05
o:	14.01	15.45
I	7.18	7.81
E	8.65	9.17
A	8.79	9.26
O	9.36	10.23
Y	8.49	8.74
@	7.98	10.11
Ei	15.49	16.00
9y	12.66	15.45
Au	15.17	17.41

Table 1: Average vowel duration when not or when under stress.

As stated in the introduction, in [5] it was reported that incorrect recognition results often have an odd segmentation, with segments that are either too long or too short. The notion 'too long/short' is a rather subjective concept, if only because the expected duration of the phones is strongly related to the average speech rate. In order to base a confidence measure on this phenomenon we have to define a metric that reflects how much the duration of individual segments deviates from the expectation based on the average speech rate of the utterance.

In the scope of this paper we will examine duration based measures on syllabic level rather than phone level. The main reason for this choice is that syllable rate is probably somewhat more stable than phone rate. In fast and conversational speech phones may be deleted in some contexts, while complete zyllable deletions are less likely. Thus, syllable based measures should be more stable than their phone based counterparts.

2.2. Speech rate factor

We define the speech rate factor of a syllable token s_i as the length (i.e. number of frames) of s_i divided by the expected length of s_i.

$$srf(s_i) = \frac{len(s_i)}{E(len(s_i))} \tag{1}$$

The expected length of s_i is computed by adding the average lengths of the phoneme constituents of s_i, which can be reliably estimated on a moderate size train set, as described in the previous section. From now on we will refer to $srf(s_i)$ with the term srf_i.

The measure for Speech Rate Confidence (SRC) of a word W with syllable speech rate factors $srf_1 \ldots srf_{nbsyl}$ can now be computed with formula (2):

$$SRC(W) = \sqrt{\sum_{nbsyl} (srf_i - m)^2 \Big/ nbsyl - 1} \tag{2}$$

where m is the mean speech rate factor in the utterance, obtained with formula (3):

$$m = \frac{1}{nbsyl} \sum_{i=1}^{nbsyl} srf_i \tag{3}$$

Actually, SRC is simply the standard deviation of the speech rate factor in the word W. Note that this formulation implies that we cannot compute the SRC for monosyllabic utterances. To solve this problem we assign a fixed SRC value to monosyllablic words, viz. the average SRC of all multisyllabic words of our development set. Additionally, we introduce a second measure that is only based on the average speech rate factor.

$$SRD(W) = |1 - m| \tag{4}$$

The *speech rate deviation* (SRD) in Formula (4) expresses how much the mean speech rate factor differs from the overall mean of the train samples (=1). Both speech rate factor measures presented above can be computed over all syllables in a word, or over the stressed and unstressed syllables separately. In the latter case phone duration statistics are computed for stressed and unstressed syllables.

2.3. Log Likelihood Ratio score

The usefulness of Log Likelihood Ratio (LLR) measures for utterance verification has extensively been demonstrated in many studies, for instance [2] and [7]. When N-best decoding is available, the ratio of the log likelihood of the best hypothesis and the runner-up hypothesis can give an indication of the confusability at sentence or word level. However, as we demonstrated in [8], this measure becomes virtually meaningless when the proportion of OoV speech grows beyond a certain threshold. Under those circumstances a hypothesis testing approach on subword level is much more effective. A survey of the development test database for the present experiment showed that the present task, city name recognition, is subject to OoV speech in 7.3% of the utterances. This proportion

motivates the use of the hypothesis testing approach of LLR. We define the frame based LLR as in formula 4:

$$LLR(X \mid \varphi) = \frac{\log P(X \mid M_{\varphi})}{\log P(X \mid M_{\overline{\varphi}})} \quad (5)$$

where X is a string of acoustic feature vectors, and φ is an hypothesized subword for the concerned frames. In the present study we have chosen to use context independent phones as our subword units. M_{φ} and $M_{\overline{\varphi}}$ are models for the target and anti-phone respectively. The anti-phones model the alternative hypothesis, viz. that phone φ was *not* realized in the observation X.

Among many ways to define the anti-phone of each target phoneme, models trained on all speech but the target have proven to be quite effective for computing confidence measures [2][9]. We will adopt this strategy, with the following reservation: phones that have virtually identical spectra as the target phone must be excluded from the training material of the anti-models. Table 2 lists the target phones (1st and 3rd columns) and phones that were excluded from the training material for the anti-models (2nd and 4th columns).

Target (= 'positive') and anti-models are HMMs, each having mixture pdfs of maximally 32 Gaussians per state. In our study we computed an LLR for every frame of the recognised utterance. Next we average all frame LLR scores on phone level. Finally we compute the average of these phone scores on word level. The study reported in [10] has shown that this two-stage averaging outperforms the simpler, direct computation of a word-level average of all frame scores. In the rest of this paper, we refer to this feature with 'LLR'.
Since we want to examine whether it helps in verification to treat phones in stressed syllables separately, we train two sets of verification models; models that do and do not distinguish between train tokens have lexical stress.

Target	Excludes (and vice versa)	Target	Excludes (and vice versa)
A	a:	@	Y
e:	I	m	n N
o:	O	n	m N
b	P	N	m n
t	D	h	@ A E I O Y a: e: i o: u Ei Au 9y 2:
f	V		
k	G		
s	Z	Z	S

Table 2: The phonemes in the target column exclude the phonemes on their right from the antimodels.

2.4. Word length

Word length, measured in the form of the number of syllables in the canonical transcription of the vocabulary items is a very simple, but potentially powerful measure of the confidence of the ASR output. Short monosyllabic words may be more confusable than long polysyllabic words, even if the task pertains only to content words that are expected to carry a pitch accent. Therefore, we include the number of syllables of the words as an additional prosodic measure. The shorthand notation for number of syllables will be SYL.

2.5. SNR

Since we conduct our experiments for a real world Directory Assistance application, a wide variety of adverse acoustic conditions can be expected. Since the training material only contains tokens with a relatively high signal-to-noise ratio (SNR), high noise levels in the test database will result in an acoustic mismatch with the models trained on clean data. This mismatch may not only affect the ASR performance, it may also interfere with the computation of confidence measures.

These considerations led us to the idea to compute SNR as a cue for the correctness of a hypothesized keyword. For this study we use a simple but fast method to estimate the SNR of an utterance. The average signal energy is computed as the root mean squared energy of 70% of all frames that contain the most energy. The average noise/background energy is computed likewise on the remaining 30% of the frames. Taking the logarithm of the ratio yields the SNR.

2.6. Combination

Having collected and computed 5 predictors (SYL, SRC, SRD, LLR and SNR) that are believed to correlate with the correctness of a recognition hypothesis, we have to define a decision strategy that uses these measures to accept or reject the utterance. Since there is no evident a priori relationship between the cues, and since we have a large number of city name utterances at our disposal, we have chosen to use a heuristic algorithm, viz. discriminant analysis. Using this method, we estimate weighting coefficients on an independent development set. These coefficients define a hyperplane in the multi-dimensional cue-space that optimally separates incorrect ASR results from the correct ones. The weighting coefficients obtained from the development set are then used to combine the five individual measures to a single confidence score for the utterances in the test set.

3. Experimental validation

To test the ideas outlined above, we set up an experiment in an ASR task for city name recognition. In order to compute the phoneme duration statistics (partly displayed in Table 1) we made an automatic segmentation of our training material, 42,101 short utterances of the Dutch Polyphone database [11] by means of forced alignment. Next we computed average duration of stressed and unstressed phones.

Target and anti-models were trained for all phoneme classes. Just like the duration statistics, we did this with and without making a distinction for lexical stress.

3.1. Corpora

The development and test material used for our experiments are subsets of the Dutch Directory Assistance Corpus (DDAC2000) [11]. The recordings are from a real nationwide directory inquiry service, in which callers were prompted to specify name of the city in which they requested a listing in isolated utterance mode. The development corpus used for the research described in this paper contains 10,954 utterances. The independent test corpus comprisses 11,499 utterances. In 97.9% of the utterances in the development corpus the caller mentions a city name or says 'I don't know'. The latter answer applies to premium rate numbers of businesses and agencies of which the modal customer does not know where they are located. 7.3% of the utterances contain at least one Out of Vocabulary (OOV) word.

Recordings were made from the public switched telephone network. The signal was sampled at 8 kHz and stored in a-law format. Acoustic pre-processing comprised extracting 14 MFCCs (c0..c13) and their first-order derivatives from 16 ms Hamming windowed frames, with a 10 ms shift.

3.2. Acoustic models

Acoustic models were trained on 42,101 short utterances of the Dutch Polyphone database [11]. The HMM set consists of 37 monophone models, one tristate noise model and two single state models: one for silence and one for garbage speech. The garbage model is trained on all speech frames of the train material. In each state acoustic variance is modelled by a mixture pdf of maximally 32 Gaussians.

We used our default topology for the phone and noise HMMs: 3 segments with left to right transitions. Each segment consists of two states with identical mixtures, one of which can be skipped. A low penalty is associated with a self-loop, a higher penalty for skipping one state. This typology assures that each phone model consumes at least 3 acoustic vectors (30 ms) of the input signal.

3.3. Lexicon

The lexicon contains all 2,377 Dutch city names, 12 province names, 3 garbage tokens of different length, 1 non-speech noise symbol, 2 entries for filled pauses, 3 multiword expressions for 'I don't know' and 4 frequently used context words.

3.4. Language model

Although city name recognition is an isolated word recognition task, the speech recognizer used in our experiments actually is a continous speech recognizer that uses probabilistic language models. Consequently, utterances that contain OoV words may be recognised as a sequence of city names. In cases like these, the very first keyword has the highest probability to be correct. Therefore, we perform a postprocessing step after recognition by discarding all words following the first occurrence of a city name or 'I don't know' expression.

To steer the process of selecting lexicon items during recognition, we trained a category bigram language model with categories for city names, province names, and context expressions, such as 'the (city) name is …'. The development database is obviously too small to train a full bigram model for this task. The within-category unigram for city names was estimated on the number of streets of each city in the Dutch zipcode book. A province name is only mentioned by the caller in the exceptional case that a city's name is not unique and needs disambiguation. The unigram distribution of each member of this category was estimated on the total number of streets of all cities with ambiguous names in that province.

3.5. Evaluation

In the remainder of this paper we report recognition performance in the form of sentence error rate (SER). A sentence is correct if the city name, 'unknown', or empty value of the recognised utterance matches the value of the reference.

In order to compare different rejection strategies, we compute the Confidence Error Rate, see formula (6), and plot the ROC curve over the whole threshold domain.

$$CER(T) = \frac{FA(T) + FR(T)}{N} \tag{6}$$

where T is the threshold on the linear combination of the five individual predictors of the confidence score, *FA(T)* and *FR(T)* are the number of false accepts and rejects at the threshold value T *CER*s are reported for several combinations of confidence predictors obtained from stepwise LDA analyses on the development test corpus.

4. Results

The baseline Sentence Error Rate, computed on the recognition result of our test corpus without rejecting any output was 16.64%. 73% of the errors are due to substitution of valid city names. Deletion and insertion errors are more or less in balance. The linear combinations of the confidence measures are all computed using the baseline recognition result. In this way, the keyphrase of an utterance gets exactly one confidence score. A stepwise Linear Discriminant Analysis performed on the development corpus yielded the following set of normalised eigenvalues for LLR, SYL, SRC, SRD, and SNR: 0.5, 0.31, 0.16, 0.014, and 0.016. Thus, it appears that SRD and SNR make no significant contribution.

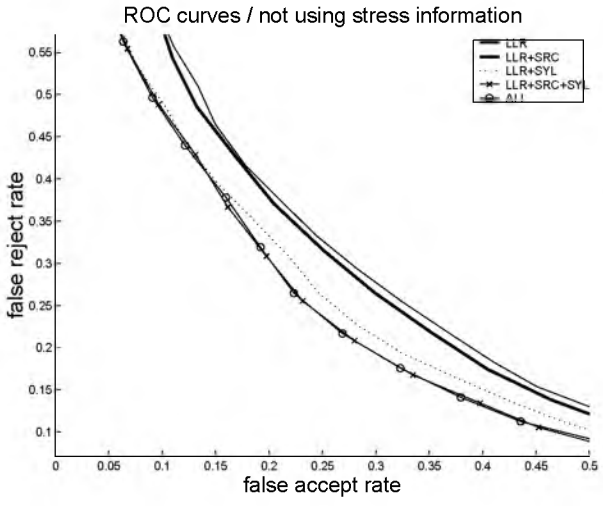

Figure 1: ROC-curves when not using stress information

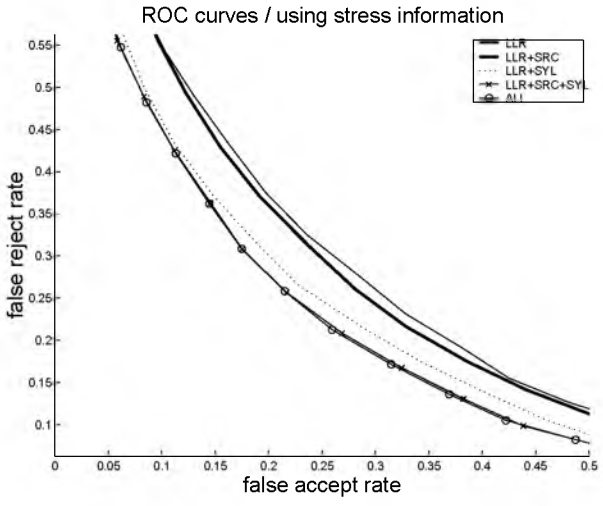

Figure 2: ROC-curves when using stress information

Table 3 displays the Confidence Error Rates for each of the combinations of confidence predictors (rows) we examined. The middle column displays CERs for confidence measures (or combinations) where no distinction was made between phones or syllables with and without lexical stress. In the rightmost column this distinction was made. For a prior probability of 16.5% error and a population size of 11,000 items the 95% confidence interval amounts to 0.6%. Therefore, the addition of SRC and SYL to LLR make for a significant improvement over LLR alone.

Combination \ Models	no stress	stress
LLR	14.7	14.7
LLR+SRC	14.3	14.0
LLR+SYL	13.9	13.4
LLR+SRC+SYL	13.7	13.1
ALL	13.6	13.1

Table 3: Confidence Error Rates for combinations of predictors (rows). In column 1 no distinction was made for whether syllables have lexical stress, while in column 2 there was.

5. Discussion

The ROC curves in Figure 1 and 2 for the conditions with and without distinction between stressed and unstressed syllables in Figure 1 and 2 show essentially identical trends; the curves are in the same order, and they have almost identical shapes. The difference between the conditions is rather in the position of the curves. The Equal Error Rates (at the intersection of the curves and the bottom-left to top-right diagonal) are all ±2% lower in the condition that separates stressed and unstressed syllables. Therefore, it appears that information on word stress can be used to advantage in the recognition of short utterances. This finding corroborates the tentavive explanation in [6] for the fact that word stress in not an effective parameter in continuous speech recognition. Apparently, word stress is informative if it is a good predictor of actual accent. Nevertheless, it is also evident that the separation of stressed and unstressed syllables does not result in a qualitatively different behaviour of the individual confidence predictors, nor of combinations thereof

Of all confidence predictors LLR is evidently the most powerful. It is interesting to see that the separation between stressed and unstressed syllables has no effect on LLR. Threfore, it seems that the spectral features of phones in stressed syllables do not differ very much from phones in unstressed syllables. The distinction between stressed and unstressed syllables does make a contribution to all duration based measures. The highest additional contribution to LLR comes from the number of syllables in the word. City names with many syllables (up to 7 in Dutch) require the speaker to produce a substantial amount of acoustic data. The higher the amount of data, the more reliable the classification decision will be. In other words, this information is directly connected to the aim of verification.

The SRC also contributes to improved separation of correct and incorrect solutions in both conditions, although its contribution to the combination of LLR and SYL may not be significant. The minor contribution of SRC may be due to the details of the way in which we accounted for within-word variation in syllable duration. Further research is required to better understand the potential contribution of segmental segmentation information to confidence measures.

SNR and SRD do not seem to add value to the first three confidence measures. This appears from the comparison of the last and penultimate row of Table 3. For SRD the same remarks apply as for SRC above. It is possible that other ways to measure the difference between actual and expected duration of words may help to make the information in total word duration more effective. However, SRD was introduced for monosyllabic words, and it is well possible that for these words duration information is overridden by the fact that they are intrinsically more difficult to recognize, a fact that is already accounted for by SYL. The fact that SRD is no correlate of correctness, may also be explained by the HMM topology. The costs associated with extraordinary values for SRD simply prohibit the Viterbi alignment to deviate too much from the average speech rate factor.

As far as SNR is concerned, it is likely that there are too few 'unacceptable' values in our (test) recordings. From previous analyses of the DDAC2000 corpus it appeared that most calls came from relatively quiet office environments. Apparently, the acoustic conditions in the test corpus did not deviate enough from the conditions in the training material to turn SNR into a useful predictor of recognition errors.

6. Conclusions

In this paper we have examined several linear combinations of prosodically motivated confidence measures for utterance verification in a city name recognition task. In one set of combinations we did not distinguish between phones in stressed and unstressed syllables, in the other we did. The results show that a combination of a measure of the number of syllables in the words and a measure of speech rate factor divergence (SRC) with a more conventional Log Likelihood Ratio (LLR) scores does profit from information about syllable stress. Of the prosodic measures investigated in this paper the number of syllables in the word proved to be most powerful.

7. Acknowledgements

This research is supported by the EC under the IST-HLT Programme. The authors thank www.spex.nl for their SNR tool.

8. References

[1] M. Weintraub, *LVCSR Log-Likelihood Ratio Scoring for Keyphrase Spotting*, Proc. ICASSP'95, Detroit, 1995, vol. I, pp. 297-300

[2] C.-H. Lee, *A Unified Statistical Hypothesis Testing Approach to Speaker Verification and Verbal Information Verification*, Proc. COST250, Rhodes, 1997, pp. 63-72

[3] C. Garcia-Mateo, W. Reichl, S. Ortmanns, *On combining Confidence Measures in HMM-based Speech Recognisers*, Proc. ASRU'99, Keystone (CO), 1999, pp. 201-204

[4] K. Bartkova, D. Jouvet, *Usefulness of Phonetic Parameters in a Rejection Procedure of an HMM based Speech Recognition System*, Proc. Eurospeech97, Rhodes, 1997, pp. 267-270

[5] S. Goronzy, K. Marasek, A. Haag, R. Kompe, *Prosodically Motivated Features for Confidence Measures*, Proc. ISCA Workshop ASR-2000, Paris, 2000

[6] D. van Kuijk, L. Boves, *Acoustic Characteristics of Lexical Stress in Continuous Telephone Speech*, Speech Communication 27 (1999), pp. 95-111

[7] D. Charlet, G. Mercier, D. Jouvet, *On Combining Confidence Measures for Improved Rejection of Incorrect Data*, Proc. of Eurospeech, Aalborg, 2001, pp. 2113-2116.

[8] G. Bouwman, J. Sturm, L. Boves, *Effects of OOV rates on Keyword Rejection Schemes,* Proc. of Eurospeech, Aalborg, 2001, pp. 2585-2588.

[9] P. Ramesh, C.-H. Lee, B.-H. Juang, *Context Dependent Anti Subword Modeling for Utterance Verification*, Proc of ICSLP '98, Sydney, vol. 7, pp. 3233-3236

[10] G. Bernardis and H. Bourlard, *Improving Posterior based Confidence Measures in HMM/ANN Speech Recognition Systems*, Proc of ICSLP '98, Sydney, vol. 3, pp. 775-778

[11] J. Sturm, H. Kamperman, L. Boves, E. den Os, *Impact of speaking style and speaking task on acoustic models*, Proc. ICSLP 2000, Beijing, 2000, pp. 361-364

[12] G. Bouwman, L. Boves, *Using Discriminative principles for Recognising City Names*, Proc. of the workshop Adaptation methods for ASR, Sophia-Antipolis, 2001, pp 109-112.